Utilización de Razonamiento Basado en Casos en el Análisis de Información Textual

D. Cordero, P. Roldán, S. Schiaffino, A. Amandi

Instituto de Sistemas ISISTAN – Facultad de Ciencias Exactas
Universidad Nacional del Centro de la Pcia. de Buenos Aires
Campus Universitario Paraje Arroyo Seco
(7000) – Tandil – Bs. As., Argentina
Email: {dcordero, proldan, sschia, amandi}@exa.unicen.edu.ar

Resumen

Encontrar información relevante para usuarios particulares en Internet es un proceso que requiere análisis de textos que pueden ser clasificados en diversas temáticas. Se expone en este artículo cómo puede aplicarse la técnica de Razonamiento Basado en Casos para la recuperación y clasificación de documentos. Nuestra propuesta está basada en la conversión de documentos textuales a una representación de casos que es utilizada para recuperar documentos similares. Se presenta también un agente inteligente como ejemplo para mostrar la aplicabilidad de esta propuesta. Este agente genera diarios digitales personales a partir del perfil de un usuario construido basándose en la observación de sus preferencias.

Palabras claves: Razonamiento Basado en Casos, Recuperación de Información.

1. Introducción

La gran cantidad de información disponible en Internet hace que en el proceso de búsqueda de documentos realizado mediante los métodos tradicionales, un usuario encuentre un amplio conjunto de páginas cuyo contenido le resulte irrelevante. El problema que surge de esta situación es el tiempo que el usuario necesita para analizar todos estos documentos, con el fin de poder distinguir cuales son de su interés.
Este costo podría ser reducido sensiblemente si se pudiera brindar al usuario la posibilidad de acceder directamente a aquellos documentos que se encuentran entre sus preferencias.

Para alcanzar este objetivo es necesario, primero, descubrir dinámicamente el tema de cada documento con el suficiente grado de detalle que permita ser distinguido de temáticas similares y, segundo, decidir si es de interés para un determinado lector.

Para poder recuperar de la Web sólo la información de interés de un usuario determinado, hemos diseñado e implementado un agente inteligente que observe la navegación realizada por un usuario a través de las distintas páginas. Registrando características particulares de estas páginas es posible clasificar los temas de lectura preferidos por el usuario. Para alcanzar el segundo objetivo, hemos desarrollado un método basado en la aplicación de Razonamiento Basado en Casos sobre experiencias de lecturas de páginas de Internet.

Para analizar la información recolectada por el agente se utiliza una base de temas estática. Es decir, cada documento en un sitio contiene información sobre un tema perteneciente a una categoría establecida previamente. Una base fija de temas sólo permite clasificar los documentos en categorías muy generales. Para realizar una categorización

puntual de un determinado documento, esta clasificación tan general resultaría insuficiente si se quieren identificar preferencias específicas.

Como documentos similares en su contenido se refieren al mismo tema, una posible forma de clasificarlos es agrupar aquellos con un alto grado de similitud. De esta forma se consigue incorporar un nivel más a la clasificación preestablecida.

El contenido de un documento puede describirse utilizando un conjunto de palabras representativas del mismo. La propuesta es extraer los sustantivos como una forma de representar el tema puntual, modificando su grado de relevancia dentro de un tema determinado a través de la modificación sistemática de sus pesos. Se utilizará Razonamiento Basado en Casos para realizar esta clasificación puntual representando el texto de una página en un caso.

Este artículo se organiza de la siguiente manera. En la Sección 2 se detalla la utilización de Razonamiento Basado en Casos para la categorización de documentos por contenido. En la Sección 3 se presenta la construcción del perfil de lectura de un usuario. En la Sección 4 se muestra un ejemplo para la aplicación de esta propuesta dentro del dominio de los diarios digitales. Se seleccionó este dominio de estudio dado que la información puntual contenida en sus páginas es altamente dinámica. Finalmente, se presentan las conclusiones.

2. Utilización de Razonamiento Basado en Casos en la categorización de documentos

Para analizar la información contenida en documentos se utiliza en principio una base de temas estática. Esta base está organizada en una estructura jerárquica. El primer nivel de esta jerarquía está conformado por las categorías identificadas en un dominio de análisis en particular. El siguiente nivel de la jerarquía lo conforman los temas generales definidos para cada categoría. Es posible en algunos casos incorporar un nivel de subtemas para cada uno de los temas. Los temas y subtemas se definen mediante un conjunto fijo de palabras.

Por ejemplo en el dominio de los diarios digitales las secciones que conforman el diario se corresponden con las categorías de la jerarquía. Considerando la sección deportes de un diario es posible identificar los temas: automovilismo, rugby, tenis, golf, futbol, ajedrez, basquet, voley, entre otros. En este contexto el tema automovilismo queda definido por palabras como: piloto, carrera, circuito, vueltas, campeonato.

Así cada documento en un sitio contiene información sobre un tema perteneciente a una categoría establecida previamente. Para determinar el tema general del documento, el texto se divide en palabras descartando artículos, conectores, preposiciones. Las palabras restantes se utilizan para identificar el tema, comparándolas con las palabras asociadas a tal tema. Es posible determinar el tema en base a las palabras que resultan de este primer análisis léxico teniendo en cuenta dos consideraciones. En primer lugar las palabras descartadas no aportan ningún tipo de información. Por otro lado, un análisis estadístico de distintos documentos mostró que existe una relación directa entre la cantidad total de palabras de una nota y la cantidad de términos obtenidos del análisis léxico. En la figura 1 se observa la relación antes mencionada en un gráfico de dispersión de una muestra de notas en el dominio de los diarios digitales.

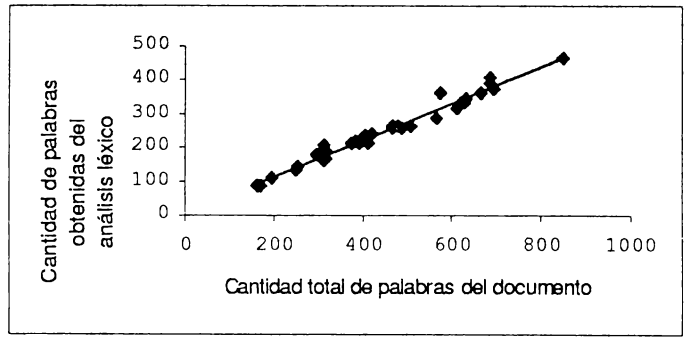

Figura 1. Relación existente entre la cantidad total de palabras y la cantidad de palabras consideradas

Para asignar el tema general se verifica que un porcentaje mínimo de las palabras se encuentren entre aquellas que definen al tema. Este porcentaje mínimo se determinó para cada tema en base a estadísticas extraídas a partir de una muestra de documentos y de la base estática de temas utilizada. En la figura 2 se muestra la distribución de frecuencias del porcentaje de palabras de un documento incluidas en el conjunto de términos que definen el tema general denominado elecciones. A partir del mismo se determinó el mínimo porcentaje a superar para que el tema de un documento del dominio de los diarios digitales sea elecciones.

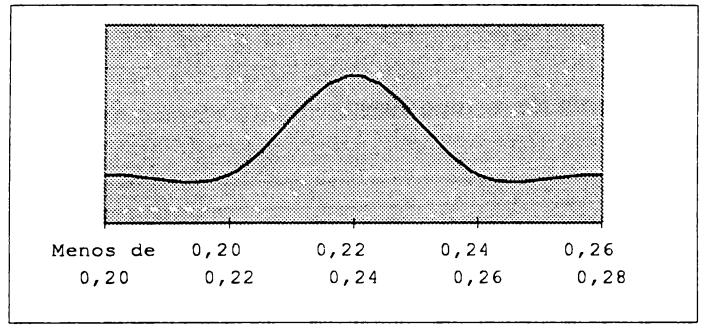

Figura 2. Distribución de la frecuencia del porcentaje de palabras incluidas en el tema elecciones

Una base fija de temas sólo permite clasificar los documentos en categorías muy generales. Si se desea realizar una categorización más fina de un determinado documento, esta clasificación posiblemente resultaría insuficiente y podría volverse obsoleta rápidamente. Esto se debe a que, en este nivel de detalle, el tipo de información que contienen los sitios digitales es dinámico, pues surgen nuevos temas continuamente. Esta característica es propia de las noticias, por lo tanto puede observarse en los diarios digitales.

Como documentos similares en su contenido se refieren a temas similares, una forma de clasificarlos es agrupar aquellos con un alto grado de similitud en sus textos, una vez que se hayan detectado los componentes textuales que los caracterizan. De esta forma se consigue incorporar un nivel más a la clasificación preestablecida.

El contenido de un documento puede describirse utilizando un conjunto de palabras representativas del mismo. Primeramente se extraen los sustantivos destacando entre ellos

los sustantivos propios pues brindan mayor información acerca del tema al que se refiere el documento, ya que representan nombres de personas, países, empresas, etc. Los sustantivos comunes se consideran palabras relevantes porque definen el contexto del tema que trata el documento.

La utilización de Razonamiento Basado en Casos permite realizar una clasificación específica haciendo uso de la información ya registrada de documentos considerados de interés para el usuario [2].

Con el fin de clasificar un documento se le asigna un tema codificado basándose en las palabras recuperadas que representan su contenido y en los documentos anteriormente clasificados. Para modelar este problema se considera que cada lectura de una nota es un caso, donde cada dimensión del caso es un par atributo - valor que registra la información relevante del documento leído.

Ante un nuevo documento a clasificar, se recuperan de la base de casos aquellos casos que representan documentos similares. Para ello se utilizan como índices la categoría a la cual pertenece el documento y el tema asignado al mismo a partir de la jerarquía de temas estática. De los casos obtenidos se selecciona mediante los procesos de matching y ranking aquel que más se asemeje a la nueva situación. La solución propuesta a tal situación es el tema puntual del caso recuperado, representado por un código. Finalmente este nuevo caso se almacena en la base de casos para ser tenido en cuenta en clasificaciones posteriores.

En el cuadro que se muestra a continuación, se presentan dos casos a ser comparados, cada uno de ellos compuesto por un conjunto de descriptores que representan la información relevante de los documentos leídos. En el dominio de los diarios digitales las dimensiones de cada caso corresponden al diario al que pertenece la página, la sección, el tema, el tema más específico, el conjunto de palabras representativas del texto, el total de palabras y el total de palabras representativas.

Caso100	**Nueva situación**
diario (La Nación)	diario (Clarín)
sección (deportes)	sección (deportes)
tema (automovilismo)	tema (automovilismo)
temaMasEspecífico (Formula Uno)	temaMasEspecífico (Formula Uno)
palabrasRelevantes	palabrasRelevantes
<Formula,2><Uno,2><Hakkinen,2>	<F,1><Uno,1><Hakkinen,7>
<McLaren,5><finlandés,2>	<McLaren,3><finlandés,1>
<carrera,2><campeonato,2>	<carrera,3><campeonato,2>
<piloto,1><Ferrari,2><camino,1>	<piloto,1><Ferrari,1>
<equipos,1><Mika,1><David,1>	<circuito,1><Mika,2><Coulthard,4>
<Coulthard,1><Alesi,2>	<vuelta,1><Michael,2><Wurz,2>
<promedio,1><ensayos,1>	<automovilismo,2><Premio,1>
<Michael,1><Schumacher,2>	<Irvine,1><neumáticos,1>
<Irvine,1><radiador,1>...	<Schumacher,1>...
CantidadTotalPalabras (424)	CantidadTotalPalabras (600)
CantidadPRs (101)	CantidadPRs (122)

Los casos son comparados unos con otros, dimensión por dimensión, teniendo en cuenta la importancia de cada dimensión en el matching. Se calcula un puntaje que representa el grado de correspondencia de un caso con otro, combinando los puntajes de matching dimensionales individuales. Los puntajes de matching total se calculan con una función de evaluación numérica que combina el grado de matching de cada dimensión con un valor que representa la importancia de la dimensión en el caso.
La función de evaluación utilizada para calcular el grado de matching total se muestra a continuación.

$$SIM(N,F) = \sum_{D_{iN} \in N,\ D_{iF} \in F} sim_i(D_{iN}, D_{iF}) * W_i$$

Donde N es el caso nuevo, F es el caso recuperado, w_i es la importancia de la dimensión i, sim_i es la función de similitud local de la dimensión i y D_i es el valor correspondiente al i-ésimo descriptor.

Como muestra la figura 3, uno de los descriptores utilizados en la representación del caso es un vector formado por pares <término,frecuencia>, el cual describe el contenido del documento. La comparación de los términos se realiza estableciendo la similitud lingüística entre los mismos. Para la comparación de frecuencias de aparición de un término en dos documentos se utilizan las tablas 1 y 2.

	Fr0	Fr1	Fr2-3	Fr>3
Fr0	0	0	0	0
Fr1	0	1	0.5	0.3
Fr2-3	0	0.5	1	0.5
Fr>3	0	0.3	0.5	1

Tabla1. Ponderación de frecuencias de aparición de la misma palabra

	Fr0	Fr1	Fr2-3	Fr>3
Fr0	0	0	0	0
Fr1	0	0.8	0.4	0.2
Fr2-3	0	0.4	0.8	0.4
Fr>3	0	0.2	0.4	0.8

Tabla2. Ponderación de frecuencias de aparición de sinónimos

El valor obtenido a partir de la función de evaluación numérica es comparado con un umbral preestablecido. Si es mayor a este umbral los casos hacen matching y al nuevo caso se le asigna el código de tema puntual que sugiere el caso recuperado. De no ser así se le agrega un código nuevo de tema puntual. Cada nueva situación genera un nuevo caso que debe ser registrado.

3. Construcción del perfil de un usuario

Un usuario habitualmente concentra su atención en temas particulares de su interés. Con el fin de presentarle sólo aquellos documentos que contienen información relevante para él, es necesario construir su perfil de preferencias.

Este perfil se obtiene a partir de la observación de los hábitos de lectura del usuario y de información que este brinda explícitamente. A partir de la observación se registra información de los documentos leídos tal como el sitio y la categoría a la que pertenece, el tema, el tiempo de lectura, entre otros. Los registros que contienen esta información se ordenan por nivel de interés. El nivel de interés queda determinado teniendo en cuenta diferentes ponderaciones: tiempo de lectura de un tema; si pertenece a los sitios, categorías o temas de interés permanentes indicados explícitamente por el usuario; respuesta del usuario ante la presentación de documentos similares en

ocasiones anteriores. A continuación se muestra la función utilizada para ponderar cada elemento p_i del perfil de un usuario, para ordenarlos posteriormente.

$$P(p_i) = \lceil (AT\ p_i / TT) * 100 \rceil + SP(p_i) + PREF(p_i) + FBK(p_i) + SS(p_i)$$

Donde AT es el tiempo acumulado de lectura del tema particular de p_i , TT es el tiempo total de navegación, SP es la función que determina la ponderación considerando si el tema puntual fue seleccionado explícitamente, PREF indica la ponderación teniendo en cuenta los intereses permanentes de lectura, FBK representa la asociación entre el tema de p_i y el feedback brindado por el usuario para él y SS es la ponderación por pertenecer a un sitio de interés permanente del usuario. Las ponderaciones obtenidas a partir de las funciones PREF y FBK pueden ser tanto positivas como negativas.
Los temas de interés que componen el perfil pueden ser tanto temas generales pertenecientes a la jerarquía estática de temas como temas codificados obtenidos utilizando Razonamiento Basado en Casos como se presentó en la Sección 2.

4. Un ejemplo de aplicación en el dominio de los diarios digitales

Para poder recuperar de la Web sólo la información de interés de un usuario determinado, hemos diseñado e implementado un agente inteligente que observe la navegación realizada por un usuario a través de las distintas páginas.

NewsAgent [3] es un agente inteligente que trabaja sobre un browser de Web tradicional, como el Netscape. NewsAgent es un tipo de agente capaz de observar al usuario a medida que éste lee las noticias en los diarios digitales y tiene la habilidad de deducir los temas de interés de un usuario en particular. Para cada usuario se construye un perfil analizando las páginas leídas por él, de la forma descripta en la Sección 3.

NewsAgent construye un diario personal para un usuario particular. Este diario personal está compuesto por notas de distintos diarios y es presentado de tal forma que los temas más interesantes se ubican en la página principal. La figura 3 muestra un diario personal el cual esta compuesto por notas de diarios escritos en español. Cada usuario puede seleccionar tanto el idioma como los diarios que el agente usará para armar el diario personal.

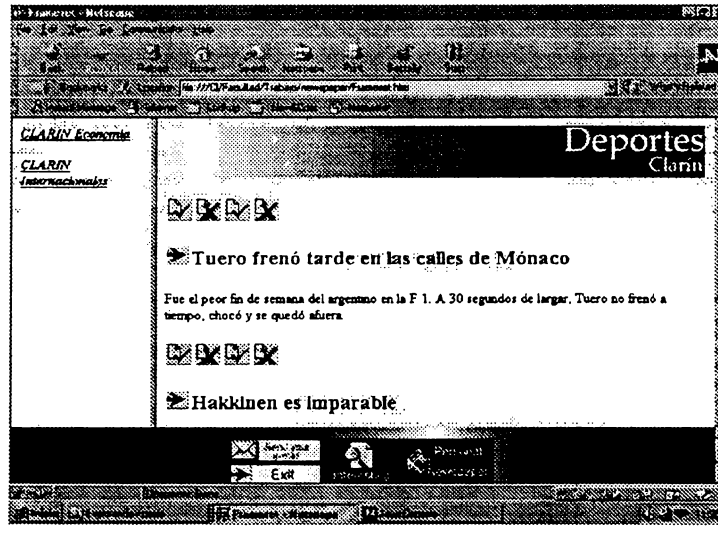

Figura 3. Diario personal generado por NewsAgent

Finalmente, NewsAgent observa al usuario mientras esta leyendo su diario personal. Con la información obtenida se modifica el estado mental del agente para la generación de los próximos diarios. Además el usuario puede indicar explícitamente si las notas presentadas fueron o no de su interés.

Conclusiones

La contribución principal de este artículo es la clasificación de documentos y la asignación de temas puntuales a los mismos, utilizando Razonamiento Basado en Casos. El agente NewsAgent está siendo utilizado por diferentes usuarios y las pruebas realizadas hasta el momento sugieren que existe una relación fuerte entre el modelo de preferencias del usuario y los diarios generados para él.

Las técnicas descriptas en este artículo serán utilizadas en el futuro para la categorización de documentos en otros dominios de interés, como documentos digitales relacionados con la agricultura, con el fin de presentarle a un usuario solamente aquellos de su interés.

Referencias

1. Kolodner J. *Case-Based Reasoning.* Morgan Kaufmann, 1993

2. Daniels J., Rissland E. *A Case-Based Approach to Inteligent Information Retrieval,* Proceedings of the SIGIR 95 Conference, Seatle 1995.

3. Cordero D., Roldán P., Schiaffino S., Amandi A. *Intelligent Agents Generating Personal Newspapers* Proceedings of the ICEIS 99, Setúbal, Portugal, Marzo,1999.